Description

An apparatus and m thod for extracting information from a formatted document

CROSS-REFERENCE TO RELATED APPLICATIONS

This is a continuation of International Application PCT/JP02/07983, published in English, with an international filing date of August 5, 2002, which claims priority to Chinese patent application 01123845.3, filed August 3, 2001, both of which are herein incorporated by reference.

Technical Field

The present invention in general relates to an apparatus and method for extracting information from an input formatted document, and in particular , to an apparatus and method for automatically extracting special character strings from an input formatted document, for example from web pages of online sale.

Background Art

It is known in the art an apparatus for extracting text information from a document, such as the technology disclosed in S. Soderland's article entitled of "Learning to Extract Text-base Information from the World Wide Web" (Proc. 3rd Intl Conf. On Knowledge Discovery and Data Mining (KDD-97)). In such an apparatus, the special character strings are distinguished by means of the character strings being the function of attribute names (e.g. "goods names") and placed before the special character strings, and are then extracted.

In the prior art apparatus, since the special character strings are distinguished and extracted by means of the character strings being the function of attribute names (such as "goods names", etc.) and placed before the special character strings, it is effective when the attribute names such as "goods names" as well as the attribute values such as "monogram accessory pouch" are available. However, the documents such as the web pages of Internet have various formats. Therefore, there is a situation that the attribute names fail to be provided. For example, only the character strings "monogram accessory pouch" are provided. In the case that the attribute names are not provided, the special character strings can not be extracted by means of the above-mentioned technology. Moreover, in the present technology, the machine can not extract the special character strings automatically, if samples are not provided manually for the machine.

Summary of the Invention

To solve the above problems, the present invention is attained. Therefore, an object of the invention is to

provide an apparatus and a method for automatically special character strings from an input formatted document.

In order to accomplish the object of the invention, there is provided an apparatus for extracting text information from an input formatted document, comprising: an input unit for inputting a formatted document; a unit for analyzing the input formatted document and saving the particular typographic information; a unit for identifying special character strings by means of the typographic information such as font size, character font, color, etc.; a unit for extracting the identified special character strings; and an output unit for outputting the extracted character strings.

According to another aspect of the invention, a method for extracting information from a formatted document is provided, which comprises the following steps: inputting a formatted document; analyzing the input formatted document and saving the particular typographic information; identifying special character strings by means of the typographic information such as font size, character font, color, etc.; extracting the identified special character strings; and outputting the extracted character strings.

According to the invention, the operations of analyzing the input formatted document, identifying special character strings by means of the typographic information such as font size, character font, color, etc and extracting the special character strings enable to automatically extract special character strings from the input formatted document and considerably increase the accuracy of extraction. Moreover, the prior apparatus requires to manually input samples for memory, while the apparatus according to the invention can automatically carry out the determination and extraction with respect to different types of the formatted document without inputting the samples.

Brief Description of the Drawings

FIG. 1 is a structural block chart of the apparatus for extracting information from a formatted document according to the invention.

FIG. 2 is document data and a flowchart illustrating a first embodiment of the invention.

FIG. 3 document data and a flowchart illustrating a second embodiment of the invention.

FIG. 4 is document data and a flowchart illustrating a third embodiment of the invention.

FIG. 5 is document data and a flowchart illustrating a fourth embodiment of the invention.

Best Mode for Carrying out the Invention

As shown in figure 1, there is a structural block chart of the apparatus for extracting information from a formatted document according to the invention.

In the extraction apparatus for extracting information from a formatted document as shown in figure 1, numeral 1 indicates an input unit for inputting a formatted document; 2 indicates a unit for analyzing the input formatted document through a certain method and saving the particular typographic information, 3 is a unit for identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc., 5 is a unit for extracting the identified special character strings, and 5 is an output unit for outputting the extracted character strings.

Next, the actions of the apparatus according to the invention will be described in detail with reference to figures 2 to 5 by an example of extracting special character strings from HTML document.

Example 1

FIG. 2 is document data and a flowchart illustrating a first embodiment of the invention, wherein figure 2 (a) is sale information which are obtained from a certain network and are a document in the form of HTML, figure 2(b) is HTML source file of the information shown in figure 2(a), figure 2(c) is a flowchart illustrating the actions of extracting information in example 1.

Next, the flow of information extraction steps in example 1 is described as follows. In step 101, HTML source file as shown in figure 2 (b) is inputted. In step 102, the thus input HTML source file is analyzed so as to find typographic information. Then, in steps 103-107, the special character strings are extracted.

At first, in step 103, the character strings to be discriminated are determined on the basis of the result obtained in step 102. Then, in step 104, a decision should be made on whether the font size of the character strings determined in step 103 is the biggest one with respect to the surrounding character strings. If it is not, then turns to the step 106. In step 106, a decision is made on whether the typographic information of said character strings is beyond the range of the preset values. If it is yes, then goes into step 107 in which the information extraction action is ended. If it is not, then returns to step 103 and thus determine the next character strings to be discriminated.

If the decision in step 104 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in

example 1 is (FONT size=5) and is the biggest among the surrounding character strings, it is determined as special typographic information. Then, goes into step 105, in which the character string "Windows Operation and Application Technology(second version)" is determined as special character strings, i.e., goods name.

Using the information extraction apparatus according to the present embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as font size.

Example 2

FIG. 3 is document data and a flowchart illustrating the second embodiment of the invention, wherein figure 3(a) is sale information which are obtained from a certain network and are a document in the form of HTML, figure 3(b) is HTML source file of the information shown in figure 3(a), figure 3(c) is a flowchart illustrating the actions of extracting information in example 2.

Next, the information extraction process in example 2 is described as follows. For clarity of illustration, the same steps as those described in the above example 1 are omitted, and only the different steps are described as below.

In step 204, a decision should be made on whether, for example, the font of the character string determined in step 203 is different from the surrounding character strings. If the decision in step 204 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in example 2 is (FONT "Chinese regular script" and the color is red(color = # ff0000)) and is particularly different from the surrounding character strings, it is determined as special typographic information. Then, goes into step 205, in which the character string "Windows Operation and Application Technology(second version)" is discriminated as special character strings, i.e., goods name.

Using the information extraction apparatus according to the present embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as font and color.

Example 3

FIG. 4 is document data and a flowchart illustrating the third embodiment of the invention, wherein figure 4(a) is sale information which are obtained from a certain network and are a document in the form of HTML, figure 4(b) is HTML source file of the information shown

in figure 4(a), figure 4(c) is a flowchart illustrating the actions of extracting information in example 3.

Next, information extraction process in example 3 is described in detail. For clarity of illustration, the same steps as those described in the above example 1 are omitted, and only the different steps are described as below.

In step 304, a decision should be made on whether, for example, the font of the character string determined in step 303 is different from the surrounding character strings. If the decision in step 304 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in this example is (FONT "Chinese regular script" and boldface (<B><FONT...</B>)) and is particularly different from the surrounding character strings, it is determined as special typographic information. Then, goes into step 305, in which the character string "Windows Operation and Application Technology(second version)" is discriminated as special character strings, i.e., goods name.

Using the information extraction apparatus according to the present embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as font and boldface.

Example 4

FIG. 5 is document data and a flowchart illustrating the fourth embodiment of the invention, wherein figure 5(a) is sale information which are obtained from a certain network and are a document in the form of HTML; figure 5(b) is HTML source file of the information shown in figure 5(a); figure 5(c) is a flowchart illustrating the actions of extracting information in example 4.

Next, information extraction process in example 4 is described in detail. For clarity of illustration, the same steps as those described in the above example 1 are omitted, and only the different steps are described as below.

In step 404, a decision should be made on whether, for example, the font of the character string determined in step 403 is different from the surrounding character strings. If the decision in step 404 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in this example is (red color (color = #ff0000) and boldface) and is particularly different from the surrounding character strings, it is determined as special typographic information. Then, goes into step

405, in which the character string "Windows Operation and Application Technology(second version)" is discriminated as special character strings, i.e., goods name.

Using the information extraction apparatus according to the this embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as color and boldface.

It should be understood, however, that the above disclosure with respect to the examples 1-4 is illustrative only, other than any limitation to the present invention. Any modifications and variations to the embodiments 1-4 of the invention may be made without departing from the spirit and the protection scope of the invention defined by the appended claims. For example, proper combination and variation of the embodiments 1-4 can be made and can obtain the same effect of the invention, i.e., automatically extracting special character strings.